REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
16-08-2012	Final Report	1-Apr-2011 - 31-Dec-2011

4. TITLE AND SUBTITLE
Dependence-Based Anomaly Detection Methodologies

5a. CONTRACT NUMBER
W911NF-11-1-0121

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Danfeng (Daphne) Yao

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Virginia Polytechnic Institute & State University
Office of Sponsored Programs
Virginia Polytechnic Institute and State University
Blacksburg, VA 24060 -

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
58056-CS-II.6

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This project addressed the fundamental problem of how to tell a system or a program is behaving properly without being compromised by stealthy malware. During the course of the project (Apr. 2011 – Dec. 2011), the PI and her students have performed studies related to designing novel dependence-based anomaly detection solutions that aim at enforcing dependence properties of legitimate programs, operations, and systems. Anomaly detection has never been systematically studied as a system security approach due to two main technical challenges: i) the (normal)

15. SUBJECT TERMS
cyber security, human centric, anomaly detection, network security, system security

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Danfeng Yao
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 540-231-8681

Report Title

Dependence-Based Anomaly Detection Methodologies

ABSTRACT

This project addressed the fundamental problem of how to tell a system or a program is behaving properly without being compromised by stealthy malware. During the course of the project (Apr. 2011 – Dec. 2011), the PI and her students have performed studies related to designing novel dependence-based anomaly detection solutions that aim at enforcing dependence properties of legitimate programs, operations, and systems. Anomaly detection has never been systematically studied as a system security approach due to two main technical challenges: i) the (normal) behaviors of legitimate programs and systems are diverse and difficult to define, and ii) unlike numerical attributes, statistical methods cannot be applied to analyzing programs and system properties; thus, there is no general enforcement methodology for normal system-security patterns. Our anomaly detection approach is to focus on enforcing the proper data and control dependencies in program execution and to identify any violations of the dependences. Such an approach yields long-lasting and powerful malware-classification solutions, because it is not limited by the constantly evolving behaviors of malware.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received	Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received	Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received		Paper
2012/08/16 2(	5	Kui Xu, Huijun Xiong, Chehai Wu, Deian Stefan, Danfeng Yao. Data-Provenance Verification For Secure Hosts, IEEE Transactions of Dependable and Secure Computing (TDSC). 9(2), 173-183.. 2012/03/01 00:00:00, . : ,
2012/08/16 2(	4	Patrick Butler, Kui Xu, Danfeng (Daphne) Yao. Quantitatively Analyzing StealthyCommunication Channels , International Conference on Applied Cryptography and Network Security (ACNS). 2011/06/07 00:00:00, . : ,
2012/08/16 2(	3	Kui Xu , Danfeng (Daphne) Yao, Qiang Ma , Alexander Crowell. Detecting Infection Onset With Behavior-Based Policies, Fifth International Conference on Network and System Security (NSS). 2011/09/10 00:00:00, . : ,
2012/08/16 2(	2	Hussain M. J. Almohri, Danfeng (Daphne) Yao, Dennis Kafura. Identifying Native Applications with High Assurance, ACM Conference on Data and Application Security and Privacy (CODASPY) . 2012/02/05 00:00:00, . : ,
2012/08/16 2(	1	William Banick, Danfeng Yao, Naren Ramakrishnan. , Hao Zhang. User Intention-Based Traffic Dependence Analysis For Anomaly Detection, Workshop on Semantics and Security (WSCS). 2012/05/24 00:00:00, . : ,

TOTAL: 5

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received	Paper

TOTAL:

Number of Manuscripts:

Books

Received	Paper

TOTAL:

Patents Submitted

SYSTEMS AND METHODS FOR THE DETECTION OF MALWARE (PCT patent filed in March 2010).

Patents Awarded

Awards

Danfeng Yao (PI) received Outstanding New Assistant Professor Award from Virginia Tech College of Engineering in 2012.

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Kui Xu	0.50	
Karim Elish	0.50	
FTE Equivalent:	**1.00**	
Total Number:	**2**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Danfeng Yao	0.10	
FTE Equivalent:	**0.10**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period:	0.00
The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:.....	0.00
The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:.....	0.00
Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):	0.00
Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:	0.00
The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense	0.00
The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:	0.00

Names of Personnel receiving masters degrees

NAME
Total Number:

Names of personnel receiving PHDs

NAME
Total Number:

Names of other research staff

NAME	PERCENT SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

This project addresses the fundamental problem of how to tell a system or a program is behaving properly without being compromised by stealthy malware. During the course of the project (Apr. 2011 – Dec. 2011), the PI and her two students (Kui Xu and Karim Elish) have performed two unique studies (user-centric dependence and quantified dependence analysis) all related to designing novel dependence -based anomaly detection solutions that aim at enforcing dependence properties of legitimate programs, operations, and systems. Anomaly detection has never been systematically studied as a system security approach due to two main technical challenges:

i) the (normal) behaviors of legitimate programs and systems are diverse and difficult to define, and
ii) unlike numerical attributes, statistical methods cannot be applied to analyzing programs and system properties; thus, there is no general enforcement methodology for normal system-security patterns.

Our anomaly detection approach is to focus on enforcing the proper data and control dependences in program execution and to identify any violations of the dependences. Such an approach yields long-lasting and powerful malware-classification solutions, because it is not limited by the constantly evolving behaviors of malware.

In user-centric data dependence, we completed stepping-stone studies on demonstrating the feasibility of simple rule-based dependence based anomaly detection in solving popular problems such as detecting drive-by downloads and the classification of Android apps. These studies provided a solid starting point for our future investigations.

We define the property of user-centric data dependence in system or program as that the system events or function calls need to be directly or indirectly in response to user actions, commands, or inputs. Despite the simplicity of this definition and the intuitive assumption on patterns of user-system interaction, we found such a data-dependence specification is sufficient for many anomaly detection needs in practice. We have two demonstrations of its usefulness in security applications.

1. DBD Detection (i.e., detecting malware-triggered file download)
We first demonstrate a concrete security application of enforcing user-centric data dependence in the context of file system access and drive-by download (DBD) detection on a host. Our work appeared in the Proceedings of Network and System Security Conference (NSS '11) [Xu 2011] and the journal version is under review at ACM TISSEC. We collected file-system events at the system call level and user-input events to a browser through keyboard and mouse hooks (in Windows), and used a rule-based decision tree for classifying where a file creation request should be allowed to happen or not. Our prototype is browser independent, and can accurately identify the benign browser-generated temporary file creations with low false positive rate. We spent significant efforts in refining and evaluating our prototype including user studies (with 21 participants) for analyzing proper threshold values, demonstrating our ability to detect 6 reproduced DBD exploits, the ability to detection 84 websites containing live DBD exploits, and automatically evaluating top 2000 (legitimate) websites ranked by Alexa.com (no false alarm
found). Select results are shown below in Figure 1 (see attachment).

2. Classification of Apps
Previous DBD study treated the program (i.e., browser) as a black box. In this study, we perform white-box program analysis for enforcing the dependence in data flow. We focus on function calls to access the critical system resources (such as network I/O, file I/O, audio interface), and inspect the dependence of their arguments on any user inputs taken by the program. Our hypothesis is that
requests to system resources in legitimate programs are typically triggered by user inputs and action, however, malware that abuses the system does not. We have developed automatic tools based on Soot (a static analysis toolkit for Java) for obtaining context sensitive data-dependence graphs. We found that in all legitimate programs, all function calls depended on user inputs, i.e., user needs to enter certain information before the request to the call is made. In most of the malicious Android apps (3 out of 4), this property of data dependence is not observed; the malicious apps abuse the system resources without user's authorization – confirming our hypothesis on the differences in user-centric data-dependence behaviors of legitimate and malicious programs. The last malware tested (Fakeneflic) is a phishing app that tricks the user to enter their Netflix login. Detecting it is out of our scope and requires site authentication (i.e., certification verification) and user education. The preliminary results are shown in Table 1 (see attachment). Our work appeared in IEEE MoST Workshop in 2012 [Elish 2012]. We are currently performing more evaluation, and plan to submit our full-version work to IEEE Security & Privacy Symposium 2013.

Summary of the most important results:
1. We demonstrated the feasibility of user-intention based dependency analysis as a general and powerful methodology for anomaly detection and system assurance.
2. We produced practical tools that can be readily used, including one for detecting DBD attacks [Xu 2011], one for classifying apps written in Java [Elish 2012], and one for identifying anomalous traffic [Zhang 2012].
3. Our other work includes a feasibility study on DNS-based botnet C&C [Butler 2011], cryptographic provenance verification for system assurance [Xu 2012], and process identification [Almohri 2012]. Please find details of these studies in the enclosed

journal/conference versions submitted.

Acknowledgments
We thank Professor Barbara G. Ryder for her advice on static program analysis. We would like to thank Army Research Office (ARO) for their support in our research investigations and our program manager Dr. Cliff Wang for his feedback on our work.

References
[Almohri 2012] Hussain M. J. Almohri, Danfeng Yao, and Dennis Kafura. Identifying Native Applications with High Assurance In Proceedings of ACM Conference on Data and Application Security and Privacy (CODASPY) . San Antonio, TX, USA. Feb. 2012. (Acceptance rate: 25%).
[Butler 2011] Patrick Butler, Kui Xu, and Danfeng Yao.
Quantitatively Analyzing Stealthy Communication Channels. In Proceedings of International Conference on Applied Cryptography and Network Security (ACNS). Lecture Notes in Computer Science. Jun. 2011 (LNCS). Acceptance rate: 18% (31/172).
[Elish 2012] Karim O. Elish, Danfeng Yao, and Barbara G. Ryder. User-Centric Dependence Analysis For Identifying Malicious Mobile Apps. In Proceedings of the Workshop on Mobile Security Technologies (MoST), in conjunction with the IEEE Symposium on Security and Privacy. San Francisco, CA. May 2012.
[Xu 2011] Kui Xu, Danfeng Yao, Qiang Ma, and Alex Crowell. Detecting Infection Onset With Behavior-Based Policies. In Proceedings of the Fifth International Conference on Network and System Security (NSS). Milan, Italy. Sep. 2011. (Acceptance rate: 22%).
[Xu 2012] Kui Xu, Huijun Xiong, Chehai Wu, Deian Stefan, and Danfeng Yao. Data-Provenance Verification For Secure Hosts. IEEE Transactions of Dependable and Secure Computing (TDSC). 9(2), 173-183. March/April 2012.
[Zhang 2012] Hao Zhang, William Banick, Danfeng Yao and Naren Ramakrishnan. User Intention-Based Traffic Dependence Analysis For Anomaly Detection. In Proceedings of Workshop on Semantics and Security (WSCS) , in conjunction with the IEEE Symposium on Security and Privacy. San Francisco, CA. May 2012.

Technology Transfer

Dependence-Based Anomaly Detection Methodologies

Final Project Report (STIR-450080)

Project Period: 9-month (Apr. 2011 – Dec. 2011)

Danfeng (Daphne) Yao
Assistant Professor
Department of Computer Science
Virginia Tech

Overview

This project addresses the fundamental problem of how to tell a system or a program is behaving properly without being compromised by stealthy malware. During the course of the project (Apr. 2011 – Dec. 2011), the PI and her students performed studies related to designing novel *dependence-based anomaly detection* solutions that aim at enforcing dependence properties of legitimate programs, operations, and systems. Anomaly detection has never been systematically studied as a system security approach due to two main technical challenges:

i) the (normal) behaviors of legitimate programs and systems are diverse and difficult to define, and
ii) unlike numerical attributes, statistical methods cannot be applied to analyzing programs and system properties; thus, there is no general enforcement methodology for normal system-security patterns.

Our anomaly detection approach is to focus on enforcing the proper *data and control dependences* in program execution and to identify any violations of the dependences. Such an approach yields long-lasting and powerful malware-classification solutions, because it is not limited by the constantly evolving behaviors of malware. For user-centric data dependence, we completed stepping-stone studies on demonstrating the feasibility of simple rule-based dependence based anomaly detection in solving popular problems such as detecting drive-by downloads and the classification of Android apps.

Report on User-Centric Dependence

We define the property of *user-centric data dependence* in system or program as that the system events or function calls need to be directly or indirectly in response to user actions, commands, or inputs. Despite the simplicity of this definition and the intuitive assumption on patterns of user-system interaction, we found such a data-dependence specification is sufficient for many anomaly detection needs in practice. We highlight some demonstrations of its usefulness in security applications.

1. DBD Detection (i.e., detecting malware-triggered file download)

We first demonstrate a concrete security application of enforcing user-centric data dependence in the context of file system access and drive-by download (DBD) detection on a host. Our work appeared in the *Proceedings of Network and System Security Conference* (NSS '11) [Xu 2011] and the journal version is under review at *Computers & Security*.

We collected file-system events at the system call level and user-input events to a browser through keyboard and mouse hooks (in Windows), and used a rule-based decision tree for classifying where a file creation request should be allowed to happen or not. Our prototype is browser-independent, and can accurately identify the benign browser-generated temporary file creations with low false positive rate. We spent significant efforts in refining and evaluating our prototype including user studies (with 21 participants) for analyzing proper threshold values, demonstrating our ability to detect 6 reproduced DBD exploits, the ability to detection 84 websites containing live DBD exploits, and automatically evaluating top 2000 (legitimate) websites ranked by Alexa.com (no false alarm found). Select results are shown below in Figure 1.

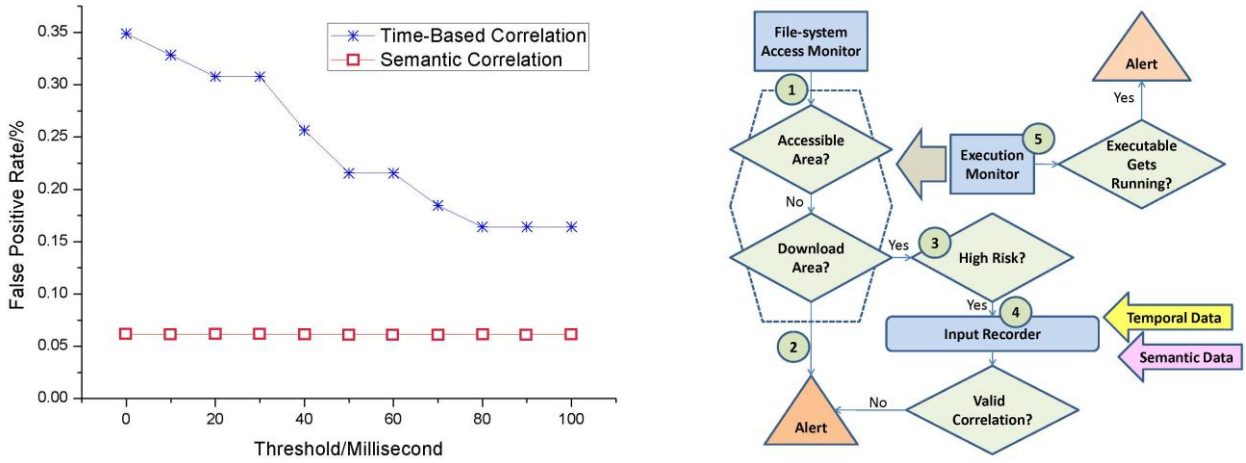

Figure 1. (a) Comparison of the false positive rates in the temporal-only dependence analysis (blue) with user data and the semantic-based dependence analysis (red) where hyperlinks associated with mouse-click events are also used in defining security rules. (b) The work flow of our prototype DeWare (standing for Deletion of Malware).

2. Classification of Apps

Previous DBD study treated the program (i.e., browser) as a black box. In this study, we perform white-box program analysis for enforcing the dependence in data flow. We focus on function calls to access the critical system resources (such as network I/O, file I/O, audio interface), and inspect the dependence of their arguments on any user inputs taken by the program. Our hypothesis is that requests to system resources in legitimate programs are typically triggered by user inputs and action, however, malware that abuses the system does not.

We have developed automatic tools based on Soot (a static analysis toolkit for Java) for obtaining context sensitive data-dependence graphs. We found that in all legitimate programs, all function calls depended on user inputs, i.e., user needs to enter certain information before the request to the call is made. In most of the malicious Android apps (3 out of 4), this property of data dependence is not observed; the malicious apps abuse the system resources without user's authorization – confirming our hypothesis on the differences in user-centric data-dependence behaviors of legitimate and malicious programs. The last malware tested (Fakeneflic) is a phishing app that tricks the user to enter their Netflix login. Detecting it is out of our scope and requires site authentication (i.e., certification verification) and user education. The preliminary results are shown in Table 1. Our work appeared in IEEE MoST workshop [Elish 2012]. We are currently performing more evaluation and formalization of the work, and plan to submit our full-version work to IEEE Security & Privacy Symposium 2013.

	Program Name	Num. of User Inputs	% of Sensitive Func. Calls without User Inputs/Sensitive Info *	Types of Function Calls
Legitimate	URLConnectionReader	1	0%	console I/O, networking
	MailSender	5	0%	javax.mail, console I/O
	UDPSendFileContent	1	0%	file I/O, networking
	SendSMS App	2	0%	telephony.GSM
Malware	EmailSpammer (proof of concept)	0	100%	javax.mail
	GGTracker.A (forwardingSMS)	0	100%	networking
	HippoSMS (sending SMS)	0	100%	telephony.GSM
	Android.Fakeneflic (Netflix)	2	0%	networking

* Number of sensitive function calls in these samples is one.

Table 1. Comparison of dependence properties in legitimate and malicious Android apps.

3. Traffic Dependency Analysis for Network Security

In this work, we investigated an approach to enforce dependencies between network traffic and user activities for

anomaly detection. We presented a framework and algorithms that analyze user actions and network events on a host according to their dependencies. Discovering these relations is useful in identifying anomalous events on a host that are caused by software flaws or malicious code. To demonstrate the feasibility of user intention based traffic dependence analysis, we implement a prototype called CR-Miner and perform extensive experimental evaluation of the accuracy, security, and efficiency of our algorithm. The results show that our algorithm can identify user intention-based traffic dependence with high accuracy (average 99.6% for 20 users) and low false alarms. Our prototype can successfully detect several pieces of HTTP-based real-world spyware. Our dependence analysis is fast with a minimal storage requirement. We give a thorough analysis on the security and robustness of the user intention-based traffic dependence approach. This work appeared in 2012 IEEE Workshop on Semantics and Security (WSCS) [Zhang 2012]. The full version of the work with expanded experiments and modeling work was submitted to ACM TISSEC.

Summary of the most important results:

1. We demonstrated the feasibility of user-intention based dependency analysis as a general and powerful methodology for anomaly detection and system assurance.
2. We produced practical tools that can be readily used, including one for detecting DBD attacks [Xu 2011], one for classifying apps written in Java [Elish 2012], and one for identifying anomalous traffic [Zhang 2012].
3. Our other work includes a feasibility study on DNS-based botnet C&C [Butler 2011], cryptographic provenance verification for system assurance [Xu 2012], and process identification [Almohri 2012]. Please find details of these studies in the enclosed journal/conference versions submitted.

Acknowledgments

We thank Professor Barbara G. Ryder for her advice on static program analysis. We would like to thank Army Research Office (ARO) for their support in our research investigations and our program manager Dr. Cliff Wang for his feedback on our work.

References

[Almohri 2012] Hussain M. J. Almohri, Danfeng Yao, and Dennis Kafura. Identifying Native Applications with High Assurance In *Proceedings of ACM Conference on Data and Application Security and Privacy (CODASPY)*. San Antonio, TX, USA. Feb. 2012. (Acceptance rate: 25%).

[Butler 2011] Patrick Butler, Kui Xu, and Danfeng Yao. Quantitatively Analyzing Stealthy Communication Channels. In *Proceedings of International Conference on Applied Cryptography and Network Security (ACNS).* Lecture Notes in Computer Science. Jun. 2011 (LNCS). Acceptance rate: 18% (31/172).

[Elish 2012] Karim O. Elish, Danfeng Yao, and Barbara G. Ryder. User-Centric Dependence Analysis For Identifying Malicious Mobile Apps. In Proceedings of the Workshop on Mobile Security Technologies (MoST), in conjunction with the IEEE Symposium on Security and Privacy. San Francisco, CA. May 2012.

[Xu 2011] Kui Xu, Danfeng Yao, Qiang Ma, and Alex Crowell. Detecting Infection Onset With Behavior-Based Policies. In *Proceedings of the Fifth International Conference on Network and System Security (NSS)*. Milan, Italy. Sep. 2011. (Acceptance rate: 22%).

[Xu 2012] Kui Xu, Huijun Xiong, Chehai Wu, Deian Stefan, and Danfeng Yao. Data-Provenance Verification For Secure Hosts. *IEEE Transactions of Dependable and Secure Computing (TDSC).* 9(2), 173-183. March/April 2012.

[Zhang 2012] Hao Zhang, William Banick, Danfeng Yao and Naren Ramakrishnan. User Intention-Based Traffic Dependence Analysis For Anomaly Detection. In *Proceedings of Workshop on Semantics and Security (WSCS)* , in conjunction with the *IEEE Symposium on Security and Privacy*. San Francisco, CA. May 2012.